Temporal and identity prediction in visual-auditory events

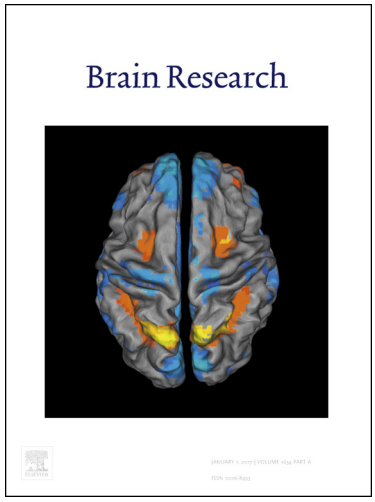
van Laarhoven, Thijs; Stekelenburg, J.J.; Vroomen, J.

Published in:
Brain Research

DOI:
10.1016/j.brainres.2017.02.014

Publication date:
2017

Document Version
Peer reviewed version



Research report

Temporal and Identity Prediction in Visual-Auditory Events: Electrophysiological Evidence from Stimulus Omissions

Thijs van Laarhoven, Jeroen J. Stekelenburg, Jean Vroomen

PII: S0006-8993(17)30076-8
DOI: http://dx.doi.org/10.1016/j.brainres.2017.02.014
Reference: BRES 45284

To appear in: *Brain Research*

Received Date: 6 October 2016
Revised Date: 13 January 2017
Accepted Date: 13 February 2017

Please cite this article as: T. van Laarhoven, J.J. Stekelenburg, J. Vroomen, Temporal and Identity Prediction in Visual-Auditory Events: Electrophysiological Evidence from Stimulus Omissions, *Brain Research* (2017), doi: http://dx.doi.org/10.1016/j.brainres.2017.02.014

Temporal and Identity Prediction in Visual-Auditory Events: Electrophysiological Evidence from Stimulus Omissions

Thijs van Laarhoven [a]

Jeroen J. Stekelenburg [a]

Jean Vroomen [a]

Tilburg University

[*]Corresponding author: T.J.T.M.vanLaarhoven@TilburgUniversity.edu, Tel: +31 13 4662994

[a] Department of Cognitive Neuropsychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Highlights

- The role of timing and identity in visual-to-auditory predictions was investigated
- Prediction was measured by the oN1 elicited by occasional omissions of the sound
- A natural audiovisual match in identity is *not* required for the auditory oN1
- The oN1 is absent when visual prediction of timing or identity is hampered
- Predictions of timing and identity are both essential elements for inducing an oN1

Abstract

A rare omission of a sound that is predictable by anticipatory visual information induces an early negative omission response (oN1) in the EEG during the period of silence

where the sound was expected. It was previously suggested that the oN1 was primarily driven by the *identity* of the anticipated sound. Here, we examined the role of *temporal* prediction in conjunction with identity prediction of the anticipated sound in the evocation of the auditory oN1. With incongruent audiovisual stimuli (a video of a handclap that is consistently combined with the sound of a car horn) we demonstrate in Experiment 1 that a natural match in identity between the visual and auditory stimulus is *not* required for inducing the oN1, and that the perceptual system can adapt predictions to unnatural stimulus events. In Experiment 2 we varied either the auditory onset (relative to the visual onset) or the identity of the sound across trials in order to hamper temporal and identity predictions. Relative to the natural stimulus with correct auditory timing and matching audiovisual identity, the oN1 was abolished when either the timing or the identity of the sound could not be predicted reliably from the video. Our study demonstrates the flexibility of the perceptual system in predictive processing (Experiment 1) and also shows that precise predictions of timing and content are both essential elements for inducing an oN1 (Experiment 2).

Keywords: stimulus omission, predictive coding, event-related potentials, visual-auditory

Abbreviations: VA (visual-auditory), MA (motor-auditory), oN1 (omission N1), oN2 (omission N2), oP3 (omission P3)

1. Introduction

One of the main and arguably most basal functions of the human brain is to 'make sense' of our environment. Understanding which events in the outside world caused activation

of specific sensory systems is what is generally considered to be the essence of perception (Lochmann & Deneve, 2011). This notion is central to the predictive coding theory, in which perceiving is considered a process of inferring the most probable causes explaining sensory signals (Friston, 2005). A key element of predictive coding is the assumption that the brain generates internal templates of the world in higher cortical areas (Mumford, 1992). These templates supposedly contain specific activation patterns of sensory systems that an occurring stimulus would normally elicit. The generated templates are presumed to be sent from higher to lower cortical processing areas (top-down), where they induce a predicted pattern of activation (Friston, 2005). If the bottom-up activation pattern induced by a stimulus matches the prediction, recognition of the stimulus occurs. Any violation of the predicted patterns by the sensory input is sent from lower sensory levels to higher cortical processing areas, reflecting the prediction error (Arnal & Giraud, 2012; Wacongne, Changeux, & Dehaene, 2012).

An approach that has been applied recently to explore the neurophysiological mechanisms of sensory prediction relies on the electrophysiological responses to infrequent unexpected stimulus *omissions*. According to the predictive coding framework, early sensory responses reflect the difference between the prediction and sensory input (Friston, 2005; Wacongne et al., 2012). During stimulus omissions there is no sensory input and the neural response to stimulus omissions is thus hypothesized to represent the neural code of top-down prediction devoid of stimulus-evoked sensory processing (Arnal & Giraud, 2012; SanMiguel, Widmann, Bendixen, Trujillo-Barreto, & Schroger, 2013b). An auditory event can be made predictable either by a motor act or anticipatory visual information regarding the onset and identity of the sound (SanMiguel et al., 2013b; Stekelenburg & Vroomen, 2015). An occasional unexpected omission of the sound evokes an early negative omission response (oN1), likely originating in the auditory cortex, suggesting that both motor and visual

predictions are able to activate a sensory template of an expected auditory stimulus in the auditory cortex.

While the available data agree that the oN1 response is an electrophysiological indicator of automatic predictive processing, it is not yet fully understood whether auditory prediction is primarily driven by *temporal* information (timing) or by the *identity* of the anticipated sound. In the motor-auditory (MA) domain, a study of SanMiguel, Saupe and Schröger (2013a) suggests that auditory omission responses are primarily driven by *identity* prediction, with only a modulatory effect of temporal prediction. In their study either a single sound or a random sound was presented after a self-paced button press. Prediction-related auditory omission responses were only observed in the single sound condition, suggesting that the sensory system, even with exact foreknowledge of the stimulus onset, does not formulate predictions if the identity of the predicted stimulus cannot be anticipated (SanMiguel, Saupe, & Schroger, 2013a). However, the timing of the sound was not specifically manipulated in their study, which calls upon further investigation of the role of temporal prediction using a stimulus omission paradigm.

The present study investigated the neural mechanisms of temporal and identity auditory predictions in the visual-auditory (VA) domain by using infrequent auditory stimulus omissions. We conducted two separate experiments. In both experiments, we used a video of an actor performing a single handclap (Figure 1) as a visual stimulus containing anticipatory information about sound identity and sound onset (Stekelenburg & Vroomen, 2007, 2015).

In the first experiment, we examined whether visual-to-auditory predictions (reflected in the omission response) are flexible and adapt, in short-term, to unnatural VA incongruences, or rather depend on long-term established associations. Compared to auditory prediction by a self-generated motor act, prediction of a sound by vision might be more affected by the informational association between the visual and auditory stimulus. While

strict informational associations are not necessarily involved in the act of a button press – as a button press can elicit various different sounds in daily practice – a video of a natural visual event may induce relatively strong auditory associations based on lifelong experience. Furthermore, although previous studies have shown that unnatural VA pairings may lead to enhancements in auditory processing (Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Thorne & Debener, 2008), it is unclear whether auditory omission responses are affected by VA congruency of identity or not. Hence, the first experiment was conducted to examine the influence of VA congruency of identity on prediction-related auditory omission responses. VA congruency was manipulated block-wise in two separate conditions. The video of the handclap was presented synchronously with either the sound of the actual handclap (natural condition) or the sound of a car horn (incongruent condition). The timing of the incongruent sound matched the timing of the natural sound. The sound of a car horn was specifically chosen to obtain a high level of VA incongruence with respect to real-world situations. VA trials were interspersed with unpredictable omissions of the sound in 12% of the trials in both conditions, c.f. SanMiguel et al. (2013a) and Stekelenburg and Vroomen (2015). Based on previous findings (SanMiguel et al., 2013b; Stekelenburg & Vroomen, 2015), three distinct omission ERP components – elicited by rare omissions of the expected sound – were expected for the natural condition: an initial negative deflection at around 50-100 ms after the expected sound onset (oN1), reflecting prediction error, followed by a second negative response at around 200 ms (oN2), and finally a more broadly distributed positive response at 300 ms (oP3), presumably reflecting higher-order error evaluation, attention orienting and subsequent updating of the forward model (Baldi & Itti, 2010; Polich, 2007). A statistically significant difference between the omission responses of the natural and incongruent conditions would suggest that the omission response depends on long-term learned VA associations.

In the second experiment, we examined the separate contributions of temporal and identity information on VA omission responses by randomizing (on a trial-to-trial basis) either auditory onset relative to visual onset or sound identity. Three experimental conditions were included: a *natural* condition, a *random-timing* condition and a *random-identity* condition (Table 1). The natural condition was identical to the natural condition of Experiment 1. In the other two conditions, either the onset (random-timing condition) or the identity (random-identity condition) of the sound was unpredictable. Temporal prediction was disrupted in the random-timing condition by presenting VA stimuli (88% of total number of trials) for which sound and vision were always asynchronous. The magnitude of asynchrony varied on a trial-to-trial basis in order to prevent adaptation to temporal asynchrony (Vroomen, Keetels, de Gelder, & Bertelson, 2004). In the random-identity condition the identity of the sound was different for each trial (c.f. the random-sound condition in SanMiguel et al. (2013a)). Based on previous findings in the MA domain, prediction-related neural activity induced by auditory omissions was expected to be most evident in the natural condition (SanMiguel et al., 2013a; Stekelenburg & Vroomen, 2015), and to be diminished in the random-identity condition (SanMiguel et al., 2013a). Assuming that timing of the sound is also of importance in the VA domain (Vroomen & Stekelenburg, 2010), we expected that the omission responses would also be diminished in the random-timing condition.

2. Results

2.1 Experiment 1

Three distinct deflections in the omission ERP were observed for both the natural and incongruent condition (Figure 2). The first negative component peaked in a time-window of 45-80 ms and is denoted as oN1. A second negative component reached its maximum at 120-240 ms (oN2). The two negative components were followed by a broadly distributed positive deflection in a window of 240-500 ms (oP3). The oN1 deflection showed a bilateral scalp

distribution with a right preponderance in both conditions, while the oN2 and oP3 components had a bilateral scalp distribution with no clear preponderance towards either hemisphere (Figure 3). Based on these scalp distributions, a left fronto-temporal (F7, F5, FT7, FC5) and right temporal (FC6, FT8, C6, T8) ROI were selected for the oN1 time-window. A frontal (F1, Fz, F2) and frontal-central (FC1, FCz, FC2) ROI was selected for the oN2 and oP3 time-window respectively. Mean amplitudes were calculated for each time-window. The presence of statistically significant omission responses was tested with separate repeated measures ANOVAs for each time-window with the within-subjects variables condition, electrode and ROI for the oN1 time-window and condition and electrode for the oN2 and oP3 time-windows.

The mean activity in the oN1, oN2 and oP3 time-windows differed from pre-stimulus baseline levels (oN1: F (1, 16) = 5.97, $p < .05$, $\eta_p^2 = .27$, oN2: F (1, 16) = 20.76, $p < .001$, $\eta_p^2 = .57$, oP3: F (1, 16) = 5.33, $p = .05$, $\eta_p^2 = .25$). Most importantly, there were no significant main effects of condition, ROI and electrode, and no interaction effects (all p values > .05), indicating that the omission responses for the natural and incongruent conditions were alike. Of note, upon visual inspection of the omission ERPs shown in Figure 2, it appears there was a difference in amplitude between the two conditions around the expected sound onset. However, statistical analysis of the mean activity recorded at the electrodes showing maximal activity in a time-window of -20-40 ms – using the same repeated measures ANOVA as used for the oN1 time-window – showed no significant main effects of condition, ROI, and electrode, and no interaction effects (all p values > .12). Figure 2 also suggests a latency difference between the two conditions in the oN2 time-window. We tested peak latency of the oN2 response in both conditions – using the same repeated measures ANOVA used for the mean activity in the oN2 time-window – and found no significant main effects of condition and electrode and no interaction effect (all p values > .19).

2.2 Experiment 2

Three distinct deflections were observed in the omission ERP of the natural condition: oN1 peaking in a temporal window of 45-100 ms; oN2 at 120-230 ms and oP3 at 240-550 ms (Figure 4). The oN1 component for the natural condition had a bilateral scalp distribution with a left preponderance, while the oN1 components for the random-timing and the random-identity condition showed a more lateralized distribution toward the left hemisphere. The oN2 and oP3 deflections had bilateral scalp topographies for all conditions. Based on these scalp potential maps (Figure 5), a left temporal (FT7, FC5, T7, C5) and right temporal (FC6, FT8, C6, T8) ROI were selected for the oN1 time-window. A frontal (F1, Fz, F2) and frontal-central (FC1, FCz, FC2) ROI was selected for the oN2 and oP3 time-window, respectively. After calculation of the mean amplitudes in each time-window, the presence of statistically significant omission responses in the oN1 time-window was tested with a repeated measures ANOVA with within-subjects variables condition, ROI and electrode. The oN2 and oP3 responses were tested with repeated measures ANOVAs with condition and electrode as within-subjects variables.

The overall mean activity in the oN1 time-window differed from pre-stimulus baseline levels, $F(1, 26) = 10.03$, $p < .01$, $\eta_p^2 = .28$. There was a main effect of condition, $F(2, 25) = 5.41$, $p < .05$, $\eta_p^2 = .30$. Post hoc paired samples t-tests (Holm-Bonferroni corrected) showed that the mean activity in the oN1 time-window was significantly more negative in the natural condition than in the random-timing condition and random-identity condition (both p values < .05). Mean activity in the oN1 time-window did not differ between the random-timing and random-identity condition. To further examine whether the oN1 differed from pre-stimulus baseline levels within each condition, we tested the mean activity in the oN1 time-window for each condition with separate repeated measures ANOVAs with within-subjects variables ROI and electrode. This analysis revealed that the mean activity in the oN1 time-window only

differed from zero in the natural condition, $F(1, 26) = 20.51$, $p < .001$, $\eta_p^2 = .44$. There were no main effects of ROI and electrode, but the ROI × electrode interaction was statistically significant, $F(1, 24) = 10.03$, $p < .01$, $\eta_p^2 = .43$. Simple effect tests examining the effect of electrode within each ROI showed no main effect of electrode in the right temporal ROI, whereas a significant main effect of electrode was revealed in the left temporal ROI, $F(3, 24) = 3.44$, $p < .05$, $\eta_p^2 = .30$. Post hoc paired samples t-tests indicated that the mean activity in the oN1 time-window was more negative at C5 than at FC5 and T7 (all p values < .05). There were no other interaction effects.

The overall mean activity in the oN2 time-window differed from pre-stimulus baseline levels, $F(1, 26) = 15.85$, $p < .001$, $\eta_p^2 = .38$. There was a main effect of condition, $F(2, 25) = 4.21$, $p < .05$, $\eta_p^2 = .25$. The mean activity in the oN2 time-window was more negative in the natural condition than in the random-timing and random-identity condition (both p values < .05). There was no difference in mean activity between the random-timing and random-identity condition. Further examination of the oN2 activity for each condition with separate repeated measures ANOVAs (with electrode as within-subjects variable) showed that the mean amplitude only differed from zero in the natural condition, $F(1, 26) = 32.08$, $p < .001$, $\eta_p^2 = .55$. There were no other main or interaction effects.

The overall mean amplitude in the oP3 time-window differed from pre-stimulus baseline levels, $F(1, 26) = 16.53$, $p < .001$, $\eta_p^2 = .39$. There was a main effect of condition, $F(2, 25) = 4.77$, $p < .05$, $\eta_p^2 = .28$. The mean activity in the oP3 time-window was more positive in the natural condition than in the random-timing and random-identity condition (both p values < .03). There was no difference in mean activity between the random-timing and random-identity condition. Testing of the oP3 activity for each condition separately – following the same procedure used on the oN2 activity – showed that the mean amplitude in

the oP3 time-window only differed from zero in the natural condition, $F(1, 26) = 16.53$, $p < .001$, $\eta_p^2 = .39$. There were no other main or interaction effects.

In sum, auditory omissions induced three distinct deflections in the natural condition: oN1 (45-100 ms), oN2 (120-230 ms) and oP3 (240-550 ms). Statistical analysis indicated that the mean activity in all time-windows differed between the natural and random-timing condition, and the natural and random-identity condition. Further examination revealed that the mean amplitude in the oN1, oN2 and oP3 time-windows, tested in the selected ROIs, only differed from pre-stimulus baseline levels in de natural condition.

3. Discussion

The current study examined the neural correlates of auditory prediction by vision using a stimulus omission paradigm. In Experiment 1 we examined whether the identity of the sound should match the natural identity of the visual information in order for the oN1 to occur, or whether an incongruent sound can also elicit the oN1, provided that the sound remains consistent across trials and synchronized with the visual event. The results of Experiment 1 showed that occasional auditory omissions in otherwise natural (video and sound of a handclap) and unnatural (video of a handclap combined with a car horn) VA combinations induced prediction-related ERP components (oN1, oN2 and oP3) of similar amplitude. This indicates that a match in identity between sound and vision of a natural event is not required per se for auditory prediction by vision. Presumably, given that the stimulus was highly predictable in both content and timing, the perceptual system learned to expect an incongruent sound, which suggests that sensory predictions adapt to unnatural stimulus events when presented repeatedly. These findings are in line with previous studies showing that unnatural VA pairings of artificial stimuli are integrated by the perceptual system in a seemingly automatic fashion (Fort et al., 2002; Giard & Peronnet, 1999; Thorne & Debener, 2008). More importantly, the current data show that visual-to-auditory predictions are not

bound to long-term established VA associations – as reflected in the highly similar omission responses in the natural and incongruent condition – but are able to adapt to unnatural VA incongruences. This ability may be crucial in order to deal with the inherent imprecision of visual to auditory predictions in real life situations.

In Experiment 2, the relative contribution of temporal and identity prediction to omission responses was explored by varying either the relative timing of the sound (while keeping sound identity constant) or the identity of the sound (while keeping relative timing constant). We found that only in the natural situation – where sound onset and identity were highly predictable from visual context – the oN1 and subsequent mid- and late latency responses (oN2, oP3) occurred. No omission responses were observed if either temporal or identity prediction was disrupted. This thus suggests that VA prediction is dependent on both *timing* and *identity*.

The results of Experiment 2 are partly consistent with studies on stimulus prediction as measured by the attenuation of the auditory N1. The amplitude of the auditory N1 is hypothesized to be a reflection of the prediction error (Arnal & Giraud, 2012; Friston, 2005). As an example, when an incoming sound matches the predicted stimulus, the amplitude of the auditory N1 is attenuated, while the neural response is enlarged when the prediction error is large. Several studies have indeed demonstrated that the amplitude of the auditory N1 is significantly attenuated when sounds are self-initiated compared to sounds triggered externally (Bass, Jacobsen, & Schroger, 2008; Martikainen, Kaneko, & Hari, 2005), or when a sound is preceded by a visual stimulus that reliably predicts the onset of the sound (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005). Our results support a study on predictive processing in the MA domain (Bass et al., 2008), which showed that attenuation of the auditory N1 depended on both the identity and timing of the auditory stimuli – with less attenuation when the auditory stimuli varied randomly in pitch and timing

relative to the motor act. In the VA domain, randomization of VA asynchrony also abolished the attenuation of the auditory N1 (Stekelenburg & Vroomen, 2007). Interestingly, though, VA congruency of identity had no effect on N1-suppression (Klucharev, Mottonen, & Sams, 2003; Stekelenburg & Vroomen, 2007). How can the latter results be reconciled with the current data showing an effect of identity on predictive processing? It could be reasoned that attenuation of the auditory N1 and the elicitation of the oN1 reflect different processes. However, an argument against this view is that the neural source of the oN1 and the attenuation of the auditory N1 induced by the same visual stimulus – the handclap video – appear to be similar (Stekelenburg & Vroomen, 2015; Vroomen & Stekelenburg, 2010), despite obvious limitations in spatial resolution of EEG. Assuming that both the oN1 and attenuation of the N1 reflect corresponding stages in predictive processing, the issue remains that different experimental paradigms produce different outcomes regarding identity predictions in the VA domain. However, a solution to this contradiction may lie in the manipulation of congruency of identity. In studies showing an effect of identity on early prediction related potentials, the incongruent trials consisted of many different incongruent VA pairings (Bass et al., 2008; SanMiguel et al., 2013a), whereas studies showing no effect of identity used only a limited number (2 to 4) of different incongruent VA pairings (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007). Considering the results of Experiment 1 of the current study, we speculate that in these latter studies participants could adapt to the violations of visual prediction of identity because a limited number of different incongruent VA pairings was repeated several times. Therefore, participants may have learned to expect either of the few different sounds. This expectation is presumably incorporated in the predictive model. Considering the numerous different incongruent pairings included in the random-identity condition of Experiment 2, no predictive model of identity could be constructed here, and hence no omission responses were elicited. In a future study it would

therefore be interesting to test whether in the VA domain N1-suppresion is diminished or abolished if multiple incongruent VA pairings were presented as in the random-identity condition of Experiment 2. Likewise, it would be of interest to examine if an omission response is induced if natural VA stimuli (allowing precise prediction of timing and identity) are presented in the context of a larger and more varied stimulus set.

The results regarding the natural and random-identity conditions of Experiment 2 are in accordance with the auditory omission study in the MA domain (SanMiguel et al., 2013a). In both studies, auditory omission responses were elicited in the natural condition but not in the random-identity condition. The new finding – besides the fact that we now tested the antecedents of predictive coding in the VA domain instead of the MA domain – is that no omission responses were elicited when a temporal prediction could not be formulated. The studies that explicitly varied auditory onset relative to visual or motor onset all agree with Experiment 2 on the importance of the timing of the to be predicted stimulus (Bass et al., 2008; Vroomen & Stekelenburg, 2010). Based on their results, SanMiguel et al. (2013a), however, did not ascribe a critical role to temporal prediction. Although the role of temporal prediction was not specifically examined in their omission study, the fact that no omission responses were observed when timing – but not identity – of the sound was predictable, led SanMiguel et al. (2013a) to conclude that motor-to-auditory prediction is primarily based on identity, with only a modulatory role for timing. The data of SanMiguel et al. (2013a) and the random-identity condition of the current study indeed suggest that identity is a prominent factor in stimulus prediction. However, if temporal prediction is indeed only of secondary importance, one would expect similar omission responses for both natural and random-timing conditions on the basis of intact identity predictions. In our opinion, the results of Experiment 2 therefore demonstrate that timing *does* play an important role in stimulus prediction, since no omission responses were observed when the timing of the sound was unpredictable.

Visual-to-auditory prediction is thus greatly hampered if the auditory onset cannot be predicted from the visual context.

The critical role of timing in predictive models fits within a theory of stimulus prediction in which the brain generates predictions of "when" parallel to "what" (Arnal & Giraud, 2012). Predictive timing ("*when*") and predictive coding ("*what*") are thought of as integral parts of a common framework, although with different functions and underlying neural bases in terms of neural rhythms. The alleged function of predictive timing is to facilitate sensory processing – taking less into account the validity of the prediction – by alignment of low frequency oscillations relative to incoming stimuli. Predictive coding concerns content-specific predictions driven by a combined role of gamma and beta oscillations. The common framework of predictive timing and coding assumes that only when an event falls inside the expected temporal window the anticipated stimulus is compared to the actual input. Our data concur with this notion of a common framework of timing and identity predictions and demonstrate that reliable prediction of the timing of the anticipated stimulus may serve as a precondition for identity prediction. Experiment 2 demonstrated that intact prediction of either solely timing or identity was insufficient to elicit prediction-related activity, thus indicating that only when the auditory stimulus is correctly timed to its anticipated onset, stimulus-specific predictions can be made. Future studies might investigate the contribution of temporal- versus identity prediction to omission responses in the MA domain by contrasting similar experimental conditions as used in the current study. It should be noted that our study cannot rule out the possibility that, conversely, intact identity prediction is necessary for temporal prediction. However, other electrophysiological studies on intersensory prediction show that visual to auditory prediction – reflected in the suppression of the auditory N1 – does not depend on audiovisual congruency of identity (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007), but is abolished when sound onset

could not be accurately predicted from the visual signal (Vroomen & Stekelenburg, 2010). This is in line with our initial interpretation of our data and favors the notion that identity prediction is more dependent on timing prediction than vice versa.

An alternative account for the results of the random-timing condition of Experiment 2 we have to consider, is that stimulus-specific predictions did remain intact, but due to the random onset of the sound the omission responses were jittered over time and smoothed out across the omission ERP. Similarly, it might also be argued that in the random-timing condition, the sensory system develops a set of predictions that corresponds to the ranges of SOAs the participant has been confronted with. In this view, participants thus may expect sounds to occur either too early (i.e., between -250 and -170 ms) or too late (i.e., between +210 and +320 ms). If confronted with an auditory omission, participants may predict by the time that the natural sound would have occurred (typically at 0 ms) that the forthcoming sound will be late – the well-known foreperiod effect, for a review see (Niemi & Naatanen, 1981). Following this reasoning, one expects the oN1 to be elicited starting at approximately 320 ms after the natural sound onset (i.e. the last possible time-point at which a sound may have occurred in the random-timing condition). However, inspection of Figure 4 shows that there was no negative deflection in the omission ERP of the random-timing condition within 200 ms after this time-point, which makes the probability of a time-jittered prediction less likely, although future studies might examine this more carefully.

The oN2 and oP3 followed the oN1 in the natural condition, but were absent when the oN1 was not elicited in the random-identity and random-timing conditions. These results mirror those of SanMiguel et al. (2013a, 2013b) in the MA domain, who also report a strict coupling between the oN1 and oN2-oP3. The strong coupling of N1, N2 and P3 components is often found in oddball paradigms (Escera & Corral, 2007), but there is also evidence that a P3 response can be elicited without a concurrent N1-N2 response (Horvath, Winkler, &

Bendixen, 2008). The oN2 is thought to reflect a higher-order error evaluation associated with stimulus conflict – in this case a conflict between the visually anticipated sound and the omitted sound. The oP3 probably reflects attention orienting triggered by the unexpected omission of the sound, and the subsequent updating of the internal forward model to minimize future error (Baldi & Itti, 2010; Polich, 2007). Assuming that the oN2 and oP3 are manifestations of processing of stimulus deviancy, the question is why they were not elicited in the random-identity and random-timing conditions? It could be reasoned that, despite the fact that the timing and identity of the sound could not be predicted in the random-timing and random-identity condition, participants still might have perceived the auditory omissions in these conditions as deviant events. Still, because of the severe disruption of the predictive model, the sensory system likely did not assign any significance to stimulus deviancy and failed to see the need for updating of the forward model because there was no viable model to be updated. The dissociation between the oN1 and oN2-oP3 in the current data suggests that stimulus prediction and stimulus deviancy are not processed in parallel, but rather points to a serial organization of different processing stages in which deviant events are only processed in depth if both timing and identity prediction can be formulated (SanMiguel et al., 2013b).

To conclude, auditory omission responses adapted to unnatural VA incongruences – such that they were highly similar to natural VA auditory omission responses – and they abolished if either the timing or the identity of the sound could not be predicted from visual context. Together, these findings suggest that predictions of timing and content are both essential elements for stimulus prediction in the VA domain.

4. Experimental Procedure

The study was conducted in accordance with the Declaration of Helsinki. All experiments were undertaken with the understanding and written consent of each participant. The Ethics Review Board of the School of Social and Behavioral Sciences of Tilburg

University approved all experimental procedures (EC-2016.48). All participants received course credits. None were diagnosed with a neurological condition and none reported use of medication. All participants reported normal hearing and normal or corrected-to-normal vision.

4.1 Experiment 1

4.1.1 Participants. Seventeen students of Tilburg University (11 female, all right-handed) with a mean age of 20.82 years (*SD* = 2.92) participated in the study.

4.1.2 Stimuli. Auditory stimuli consisted of recordings of a handclap and a car horn of equal length (200 ms) and sampling rate (44.1 kHz), with matched amplitudes based on the root mean square method. Audio files were presented over JAMO S100 stereo speakers, located directly on the left and right side of the monitor, at approximately 61dB(A) sound pressure level. Visual stimuli consisted of a video recording portraying the motion of a single handclap (Figure 1). The video started with the hands separated. Subsequently, the hands moved to each other and after collision returned to their original starting position. Total duration of the video was 1300 ms. The video was presented at a frame rate of 25 frames/s on a 19-inch Iiyama Vision Master Pro 454 CRT monitor at a refresh rate of 100 Hz, a resolution of 640 x 480 pixels (14° horizontal and 12° vertical visual angle) and at a viewing distance of approximately 70 cm.

4.1.3 Procedure. Participants were individually tested in a sound attenuated and dimly lit booth. They were instructed to carefully listen to the presented audio files and to maintain their focus on the center of the screen. VA congruency was manipulated block-wise in two separate conditions: a natural condition and an incongruent condition (Table 1). During VA trials in the natural condition, a video of a handclap was presented synchronously with the audio recording of the actual handclap. For the incongruent condition the handclap sound was replaced by the sound of a car horn. This sound was specifically chosen to obtain a high level

of VA incongruence with respect to real-world situations. In both conditions, the sound occurred 500 ms after video onset and 360 ms after the start of the hand movement, while the inter-stimulus interval (from auditory onset) was 1300 ms (Figure 1). VA trials were interspersed with unpredictable omissions of the sound in 12% of the trials in both conditions, c.f. SanMiguel et al. (2013a) and Stekelenburg and Vroomen (2015). These omission trials were randomly intermixed with VA trials with the restrictions that the first five trials of each block and the two trials immediately following an omission trial were always VA trials. Each condition was presented in seven blocks of 200 trials (with a short break between blocks). This resulted in a total of 1400 stimulus presentations in each condition, including 168 auditory stimulus omissions. Block order was varied quasi-randomly. After every fourth block a short block of 100 visual-only trials was presented (i.e. three visual-only blocks for each participant), during which only the visual recording of a handclap was presented. The visual-only (V) condition was introduced to correct for visual activity in the auditory omission trials (see 'EEG recording'). An auditory-only condition was not included, since a previous study using the same VA stimuli and a similar inter-stimulus interval demonstrated that unexpected omissions of the sound *as such* do not evoke a significant neural response (Stekelenburg & Vroomen, 2015). To ensure that participants watched the visual stimuli, 8% of all VA and V trials consisted of catch trials. Participants were required to respond with a button press after onset of a catch stimulus (i.e. a small white square superimposed on the handclap video, presented at the center of the screen, measuring 1° horizontal and 1° vertical visual angle). To prevent possible interference of (delayed) motor responses, these catch trials never preceded an omission trial. Participants were unaware of the total amount of catch trials presented in each block. After each block, the percentage of missed catch trials and false alarms was displayed at the center of the screen. Average percentage of detected catch trials across conditions was high (M = 98.76, SD = 1.76) and did not differ between conditions or

subjects and there was no condition × subject interaction effect, indicating that the participants attentively watched the video in all conditions.

4.1.4 EEG recording. The EEG was sampled at 512 Hz from 64 locations using active Ag-AgCl electrodes (BioSemi, Amsterdam, the Netherlands) mounted in an elastic cap and two mastoid electrodes. Electrodes were placed in accordance with the extended International 10-20 system. Two additional electrodes served as reference (Common Mode Sense active electrode) and ground (Driven Right Leg passive electrode). EEG was referenced offline to an average of left and right mastoids and band-pass filtered (1-30 Hz, 24 dB/octave). To facilitate a more direct comparison between the current data and the results of the previous auditory omission study in the MA domain, the same high-pass filter settings were applied as in SanMiguel et al. (2013a). Furthermore, the relatively long interval between visual and auditory stimulus onset might elicit anticipatory slow waves that may contaminate early ERP components (Teder-Salejarvi, McDonald, Di Russo, & Hillyard, 2002). To prevent this anticipatory activity from contaminating or simply obscuring the oN1 component, a relatively high high-pass filter of 1 Hz was applied. The (residual) 50 Hz interference was removed by a 50 Hz notch filter. Raw data were segmented into epochs of 1000 ms, including a 200-ms pre-stimulus baseline period. Epochs were time-locked to the expected sound onset in the natural and incongruent conditions, and to the corresponding timestamp in the V condition. After EOG correction (Gratton, Coles, & Donchin, 1983), epochs with an amplitude change exceeding ±120 μV at any EEG channel were rejected and subsequently averaged and baseline corrected for each condition separately. On average 6.34 percent (*SD* = 6.84) of the omission trials were rejected. There was no significant difference in rejected omission trials between conditions. The ERP of the V condition was subtracted from the VA omission ERPs of the natural and incongruent conditions to correct for the contribution of visual activity to the omission ERPs (Figure 6). Consequently, the VA-V difference waves

reflect prediction related activity – induced by unexpected auditory omissions – devoid of visual related activity (Stekelenburg & Vroomen, 2015).

4.2 Experiment 2

4.2.1 Participants. Twenty-seven students of Tilburg University (23 female, 4 left-handed) with a mean age of 19.93 years (*SD* = 2.40) participated after given written informed consent. None of them participated in Experiment 1. None reported use of prescription drugs or were diagnosed with a neurological disorder. All participants reported normal hearing and normal or corrected-to-normal vision and received credits in hours as part of a curricular requirement.

4.2.2 Stimuli and procedure. Three experimental conditions were included: a *natural* condition, a *random-timing* condition and a *random-identity* condition (Table 1). The natural condition was identical to the natural condition of Experiment 1. In the random-timing condition, the sound could either precede or follow the visual collision of the two hands at an unpredictable stimulus onset asynchrony (SOA). Based on the results of a simultaneity judgment (SJ) task ran prior to the EEG experiment (Figure 7), SOAs of -250, -230, -210, -190, -170, 210, 240, 260, 290 and 320 were chosen (all values in ms, negative and positive values indicate sound leading and following the natural synchrony point, respectively). In the random-identity condition, 100 different environmental sounds (e.g. a doorbell, barking dog or a car horn) of equal length (200 ms) and matched amplitudes were used. The video showed the same handclap as before and was presented synchronously with an environmental sound that was randomly selected in every trial out of the pool of 100 sounds (c.f. the random-sound condition in SanMiguel et al. (2013a)). The experimental design of the three conditions was identical to Experiment 1 (a total of 1400 trials per condition; 12% auditory omission trials; 8% catch trials). Each condition was presented in seven blocks of 200 trials, while quasi-random block sequences were allocated to each participant using a counterbalanced measures

design. After every sixth block a short block of 100 V trials was presented. Average percentage of detected catch trials across conditions was high (M = 98.48, SD = 2.51). There were no main effects of condition and subject and no condition × subject interaction effect, indicating that all participants attentively watched the video in each condition.

4.2.3 EEG recording. EEG recording and filtering was equivalent to Experiment 1. Epochs of 1000 ms (including a 200-ms pre-stimulus baseline period) were time-locked to the expected sound onset in the natural and random-identity conditions, and to the corresponding timestamp in the random-timing and V condition. All omission trials not rejected due to artifacts were included in the average visual-corrected omission-ERP for each condition. On average, 5.51% (SD = 5.74) of all omission trials were rejected. There were no significant differences in rejected omission trials between conditions.

Conflict of Interest Statement and Funding

The authors declare that they have no conflict of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn Sci, 16*(7), 390-398. doi: 10.1016/j.tics.2012.05.003

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw, 23*(5), 649-666. doi: 10.1016/j.neunet.2009.12.007

Bass, P., Jacobsen, T., & Schroger, E. (2008). Suppression of the auditory N1 event-related potential component with unpredictable self-initiated tones: evidence for internal forward models with dynamic stimulation. *Int J Psychophysiol, 70*(2), 137-143. doi: 10.1016/j.ijpsycho.2008.06.005

Escera, C., & Corral, M. J. (2007). Role of mismatch negativity and novelty-P3 in involuntary auditory attention. *Journal of Psychophysiology, 21*(3-4), 251-264. doi: 10.1027/0269-8803.21.34.251

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Early auditory-visual interactions in human cortex during nonredundant target identification. *Brain Res Cogn Brain Res, 14*(1), 20-30. doi: 10.1016/S0926-6410(02)00058-7

Friston, K. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci, 360*(1456), 815-836. doi: 10.1098/rstb.2005.1622

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J Cogn Neurosci, 11*(5), 473-490. doi: 10.1162/089892999563544

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr Clin Neurophysiol, 55*(4), 468-484. doi: 10.1016/0013-4694(83)90135-9

Horvath, J., Winkler, I., & Bendixen, A. (2008). Do N1/MMN, P3a, and RON form a strongly coupled chain reflecting the three stages of auditory distraction? *Biol Psychol, 79*(2), 139-147. doi: 10.1016/j.biopsycho.2008.04.001

Klucharev, V., Mottonen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res Cogn Brain Res, 18*(1), 65-75. doi: 10.1016/j.cogbrainres.2003.09.004

Lochmann, T., & Deneve, S. (2011). Neural processing as causal inference. *Curr Opin Neurobiol, 21*(5), 774-781. doi: 10.1016/j.conb.2011.05.018

Martikainen, M. H., Kaneko, K., & Hari, R. (2005). Suppressed responses to self-triggered sounds in the human auditory cortex. *Cereb Cortex, 15*(3), 299-302. doi: 10.1093/cercor/bhh131

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern, 66*(3), 241-251. doi: 10.1007/BF00202389

Niemi, P., & Naatanen, R. (1981). Foreperiod and Simple Reaction-Time. *Psychological Bulletin, 89*(1), 133-162. doi: 10.1037//0033-2909.89.1.133

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin Neurophysiol, 118*(10), 2128-2148. doi: 10.1016/j.clinph.2007.04.019

SanMiguel, I., Saupe, K., & Schroger, E. (2013a). I know what is missing here: electrophysiological prediction error signals elicited by omissions of predicted "what" but not "when". *Front Hum Neurosci, 7*, 407. doi: 10.3389/fnhum.2013.00407

SanMiguel, I., Widmann, A., Bendixen, A., Trujillo-Barreto, N., & Schroger, E. (2013b). Hearing silences: human auditory processing relies on preactivation of sound-specific brain activity patterns. *J Neurosci, 33*(20), 8633-8639. doi: 10.1523/JNEUROSCI.5821-12.2013

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J Cogn Neurosci, 19*(12), 1964-1973. doi: 10.1162/jocn.2007.19.12.1964

Stekelenburg, J. J., & Vroomen, J. (2015). Predictive coding of visual-auditory and motor-auditory events: An electrophysiological study. *Brain Res, 1626*, 88-96. doi: 10.1016/j.brainres.2015.01.036

Teder-Salejarvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Brain Res Cogn Brain Res, 14*(1), 106-114. doi: 10.1016/S0926-6410(02)00065-4

Thorne, J. D., & Debener, S. (2008). Irrelevant visual stimuli improve auditory task performance. *Neuroreport, 19*(5), 553-557. doi: 10.1097/WNR.0b013e3282f8b1b6

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A, 102*(4), 1181-1186. doi: 10.1073/pnas.0408949102

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Brain Res Cogn Brain Res, 22*(1), 32-35. doi: 10.1016/j.cogbrainres.2004.07.003

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J Cogn Neurosci, 22*(7), 1583-1596. doi: 10.1162/jocn.2009.21308

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *J Neurosci, 32*(11), 3665-3678. doi: 10.1523/JNEUROSCI.5003-11.2012

Figure captions

Figure 1. Time-course of the video used in all experimental conditions administered in Experiment 1 and Experiment 2.

Figure 2. Direct comparison of the grand average omission-ERPs between the natural and incongruent condition. Omission responses were corrected for visual activity via subtraction of the visual-only waveform and collapsed over electrodes in each region of interest (ROI). The first negative component peaked in a time-window of 45-80 ms (oN1). A second negative component reached its maximum in 120-240 ms (oN2). The two negative components were followed by late positive potentials in a time-window of 240-500 ms (oP3).

Figure 3. Scalp potential maps of the grand average visual-corrected omission responses for the natural and incongruent condition in the denoted oN1 (45-80 ms), oN2 (120-240 ms) and oP3 (240-500 ms) time-windows.

Figure 4. Direct comparison of the grand average omission-ERPs between the natural, random-timing and random-identity condition. Omission responses were corrected for visual activity via subtraction of the visual-only waveform and collapsed over electrodes in each region of interest (ROI). The first negative component peaked in a time-window of 45-100 ms (oN1). A second negative component was observed in a time-window of 120-230ms (oN2). The two negative components were followed by a positive deflection that reached its maximum at 240-550 ms (oP3).

Figure 5. Scalp potential maps of the grand average visual-corrected omission responses for the natural, random-timing and random-identity condition in the denoted oN1 (45-100 ms), oN2 (120-230 ms) and oP3 (240-550 ms) time-windows.

Figure 6. Comparison between visual-uncorrected and visual-corrected omission ERPs and oN1 scalp potential maps in the natural condition. The ERP of the visual-only (V) condition was subtracted from the visual-auditory (VA) omission ERP to correct for the contribution of visual activity to the omission ERPs. Consequently, the VA-V difference wave reflects prediction related activity devoid of visual activity.

Figure 7. Grand average percentages of simultaneity judgment (SJ) as a function of stimulus onset asynchrony (SOA). Two logistic curves were fitted to the data - one on the auditory-leading side and one on the visual-leading side. The just noticeable difference (JND) of 70% was calculated for both logistic curves and used as a reference for the smallest SOAs included in the random-timing condition of Experiment 2. This ensured that none of the visual-auditory (VA) trials in this condition would be perceived as synchronous VA events. The remaining SOAs were obtained from both curves by calculating the intersections at 60, 50, 40 and 30 percent of simultaneity judgment. Consequently, the following SOAs were included: -250, -230, -210, -190, -170, 210, 240, 260, 290 and 320 (all values in ms, negative and positive values indicate sound leading and following the natural synchrony point, respectively).

Figure 1

Figure 2 (revised)

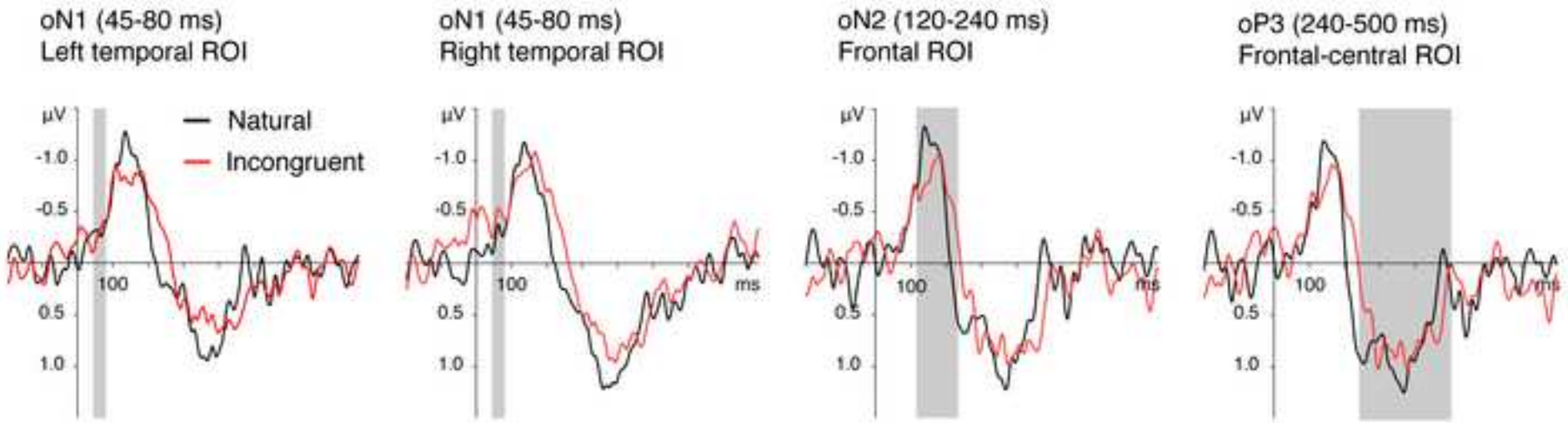


Figure 3

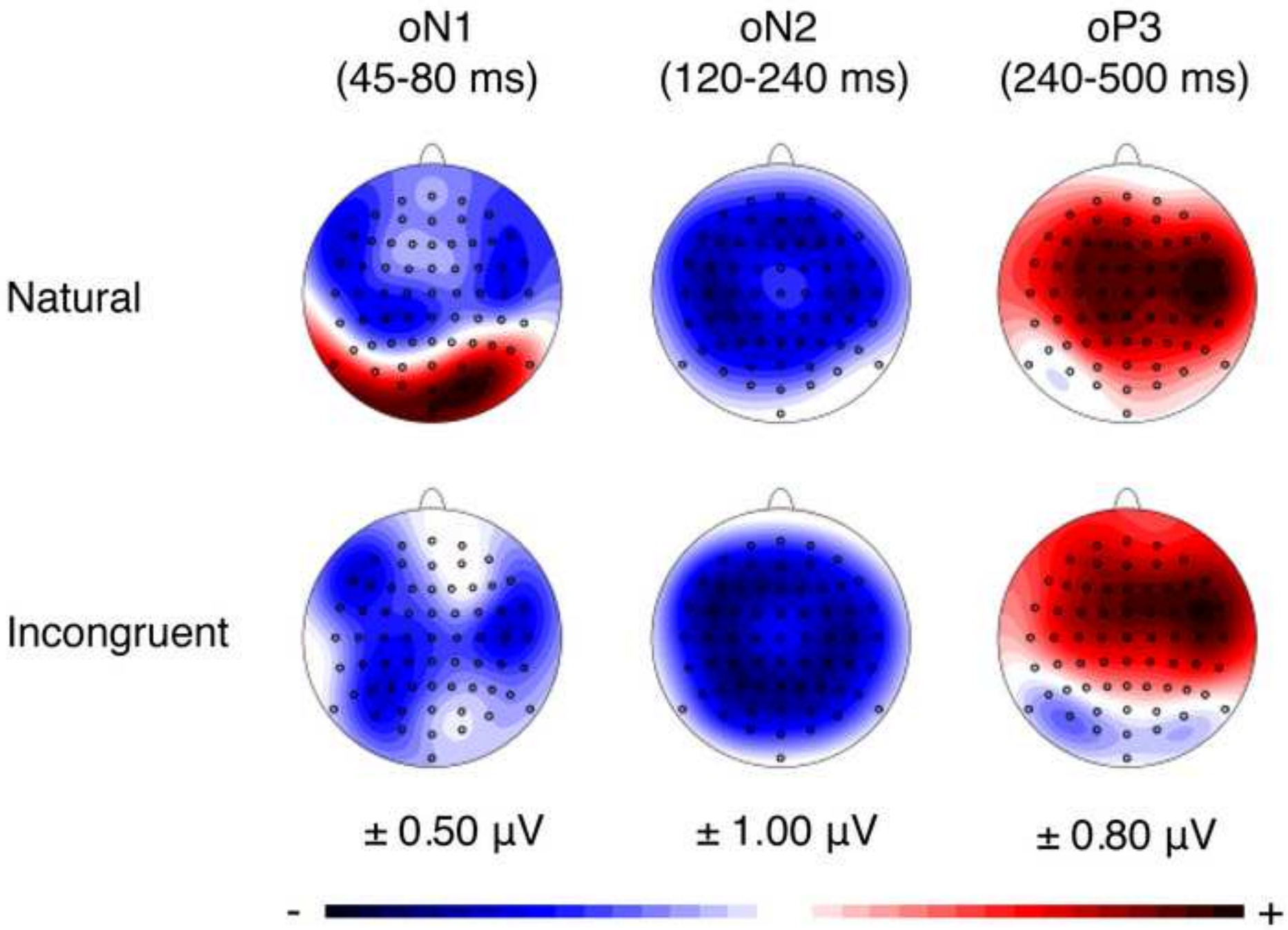


Figure 4

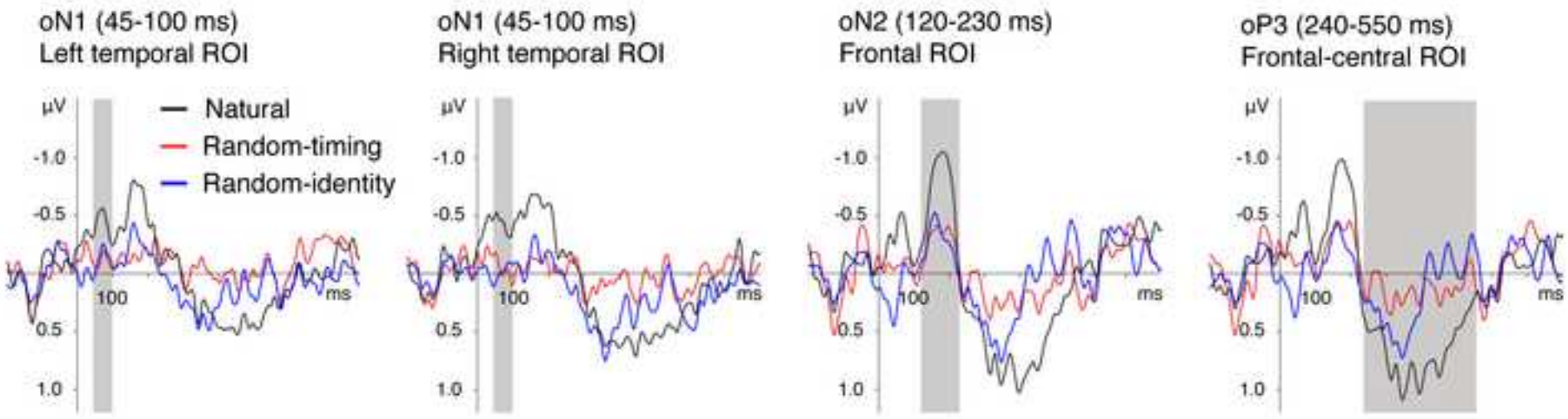


Figure 5

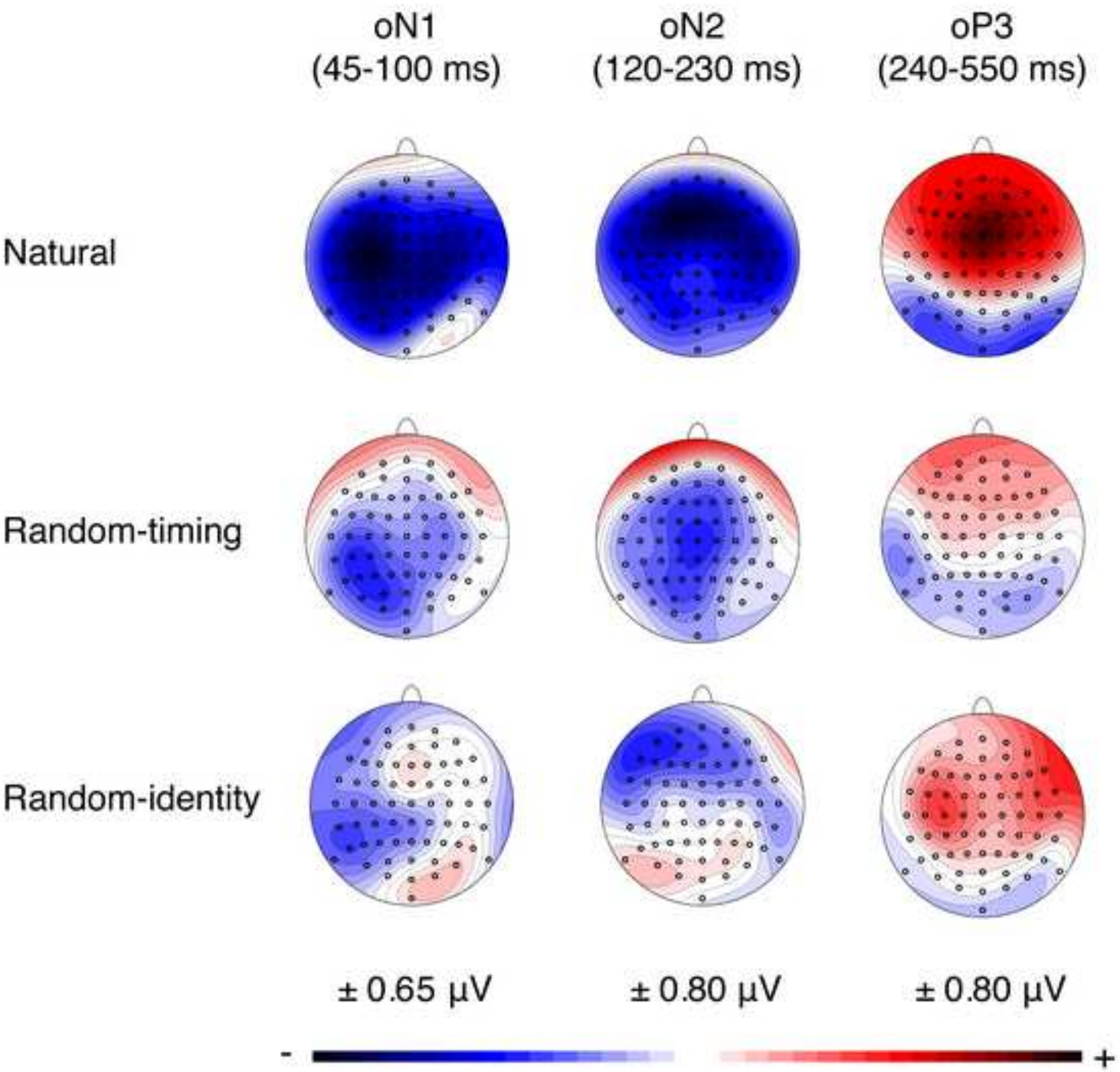


Figure 6

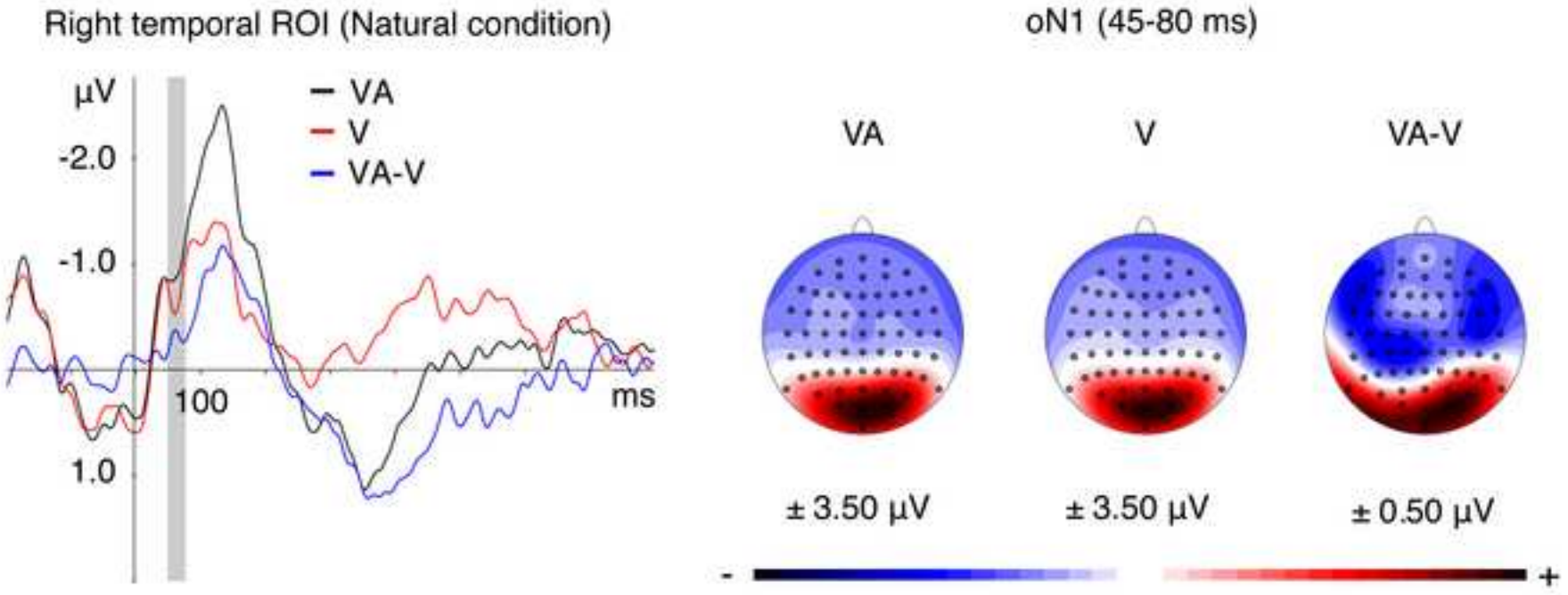


Figure 7

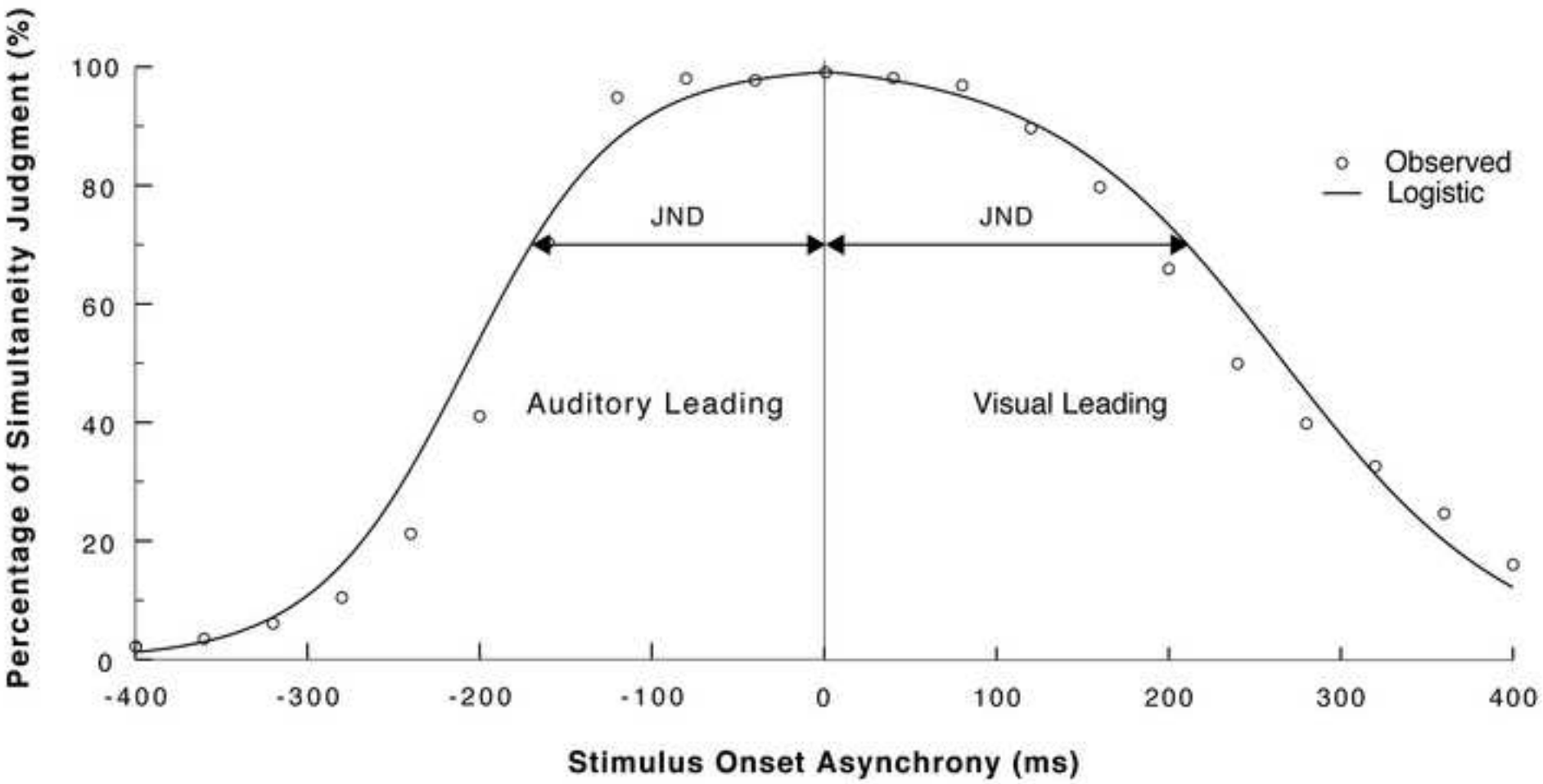


Table 1

Experimental conditions included in Experiment 1 and Experiment 2

Condition	Sound timing	Sound identity
Natural [Exp1, Exp2]	Synchronized with video	Handclap
Incongruent [Exp1]	Synchronized with video	Car horn
Random-identity [Exp2]	Synchronized with video	Random[a]
Random-timing [Exp2]	Random[b]	Handclap

[a] The identity of the sound was randomly selected in every trial out of 100 different environmental sounds (e.g. a doorbell, barking dog or a car horn) of equal length and matched amplitudes
[b] The sound could either precede or follow the visual collision moment of the two hands at a randomly selected SOA of -250, -230, -210, -190, -170, 210, 240, 260, 290, or 320 (all values in ms, negative and positive values indicate sound leading and following the natural synchrony point, respectively)